Jones, D., Matthews, J., Xie, Y., Gopsill, J., Dotter, M., & Hicks, B. (2017). Improving engineering information retrieval by combining TD-IDF and product structure classification. In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21-25.08.2017: Design Information and Knowledge, Vancouver, Canada, 21-25.08.2017 (Vol. 6, pp. 041-050)* (Vol. 6, pp. 041-050). (Proceedings of the 21st International Conference on Engineering Design ; Vol. 6). Glasgow, Scotland.

IMPROVING ENGINEERING INFORMATION RETRIEVAL BY COMBINING TD-IDF AND PRODUCT STRUCTURE CLASSIFICATION

Jones, David (1); Matthews, Jason (2); Xie, Yifan (3); Gopsill, James (1); Dotter, Martin (3); Chanchevrier, Nicolas (3); Hicks, Ben (1)
1: University of Bristol, United Kingdom; 2: University of the West of England, United Kingdom; 3: Airbus Group, United Kingdom

Abstract

Engineering Information Management (EIM) and Information Retrieval (IR) systems are central to the day to day running of large engineering organisations. The capture, interrogation, retrieval and presentation of information from design to disposal is considered to be a key enabler for greater efficiency and decision making and in turn improved productivity, profitability and competitiveness. This paper presents a contribution to the field of engineering IR through combining TF-IDF with classification against the product structure. The results of this initial investigation show that Precision, Recall and F1-Scores can be improved depending on the method of results integration and thus tailored to the search system and context.

Keywords: Knowledge management, Information management, Design informatics, Product Lifecycle Management (PLM)

Contact:
David Edward Jones
University of Bristol
Mechanical Engineering
United Kingdom
dj13730@bristol.ac.uk

Please cite this paper as:
Surnames, Initials: *Title of paper*. In: Proceedings of the 21st International Conference on Engineering Design (ICED17), Vol. 6: Design Information and Knowledge, Vancouver, Canada, 21.-25.08.2017.

1 INTRODUCTION

Engineering Information Management (EIM) and Information Retrieval (IR) systems are central to the day-to-day running of large engineering organisations. Like the Business Intelligence (BI) drive (Chen et al. 2012), the reuse of information- and data-driven management is considered a route to greater efficiency and decision making resulting in improved productivity, profitability and competitiveness (Hicks et al. 2002).

Similar to BI, a key undertaking in EIM is the development of Information Systems (IS) that better support the management of information generated across the product lifecycle. This includes the capture, interrogation, retrieval and presentation of information from design to disposal. Systems such as Product Data Management (PDM) (Liu & Xu 2001; Lee et al. 2008) and Building Information Management (BIM) (Eastman et al. 2011) have been developed to synchronise product related documentation with their respective digital models. These systems can be viewed as an amalgamation of pre-existing tools and techniques from the fields of Computer Aided Design (CAD), IR and Internet systems.

In relation to IR, one cannot disregard the continuing work and speed of development occurring across Internet technologies to further improve the indexing, searching and retrieval of online documents. Over recent years there has been a drive to improve search results by supplementing the search query with context through techniques such as personalised search or the use of Ontologies, Taxonomies and Semantics (Klampanos 2009). While both CAD and IR are mature technologies in their own rights; in terms of document search PDM, BIM and such data management systems are still found lacking in their ability to return accurate and relevant search results (Eastman et al. 2011; Stocker et al. 2014; Hawking 2004).

PDM/BIM research has begun to explore the potential in providing context to both expand and improve search results. An example being where the authors achieve an improvement using user personal preferences (Finkelstein et al. 2001). The domain specific nature of engineering and the commonality of engineering tasks present an opportunity to implement domain specific solutions. For example, (Jones, Xie, et al. 2015) discusses a classification of enterprise search queries and revealed that search queries within a large engineering organisation could be classified into ten business related classes, one of these classes being the product itself. It then stands to reason that supporting search using the highly-structured engineering product model could improve search results. However, the best method(s) for achieving this is an ongoing research challenge.

This paper presents a contribution to this research challenge through consideration of the product structure as an Ontology, where concepts are comprised of components and subsystems. Ontology expanded search has been shown to improve engineering search (Xie et al. 2011) however this article discusses a new perspective on Ontology search to realise further improvements. Term Frequency - Inverse Document Frequency (TF-IDF) is widely accepted to improve the performance of search systems yet does not always return all relevant documents due to the ambiguity of language. The work presented here combines TF-IDF and a product structure based classification system to improve the precision and recall of an IR system. The challenges of precision and recall become ever more pertinent when seeking to automatically classify documents which is one of the longer term aims of the research. That is, users rely on the classification and lose faith if results are irrelevant and get frustrated if too many results are returned.

The paper begins by discussing TF-IDF and the proposition of classifying documents against the product before expanding on the detail of how the approach was constructed, tested and evaluated.

2 BACKGROUND

There are then two aspects to the approach examined here: a TF-IDF search engine, and a classification system that classifies documents against the product structure. This section examines each of these in turn before discussing the measures of Precision, Recall and F1 Score - measures that are frequently used to compare IR techniques.

2.1 TF-IDF Search Engines

Term Frequency - Inverse Document Frequency (TF-IDF) is a corpus linguistics approach for measuring the importance of terms within a corpus (*D*) and is widely used throughout IR and machine learning

systems (Klampanos 2009). It is a combination of two measures, the *Term Frequency* (*tf*) and the *Inverse Document Frequency* (*idf*). The term frequency f is a count of the occurrence of term *t* in document *d* (Equation 1). The inverse document frequency is the natural log of the total number of documents (*N*) divided by the number of documents containing the term *t* (Equation 2). The TF-IDF is the multiplication of the two measures (Equation 3).

$$tf(t,d) = f_{t,d} \tag{1}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

$$tf\text{–}idf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

2.2 BOM Classification

Classification is one of the cornerstones of machine learning and used to label a dataset against a pre-set list of classes (Witten & Frank 2005). The benefits of classification in the context of the work presented here is that the BOM components can be used as classes against which a corpus can be classified.

Attempts to make improvements in the field or IR have examined techniques like Synsets and Ontologies/Taxonomies. These try to improve search by using relationships and concepts between terms. Synsets expand the search query by including the terms with the same meaning, for example '*powertrain*' and '*engine*'. Ontologies/Taxonomies capture the relationships between terms and concepts and use these to expand and filter searches. Continuing the example, a search for '*powertrain*' could be expanded to return the results for all the powertrain components.

Another notable field of research in this area is that of Extended or Annotated CAD that merge the CAD models and product related information (Camba et al. 2014). These systems are part of a drive to place the product and product structure at the heart of the product lifecycle. Recent work by (Jones, Chanchevrier, et al. 2015) expands on this by the suggestion of placing the product structure at the heart of engineering search. Part of the justification being that a large proportion of search within an engineering organisation are product related (Jones, Xie, et al. 2015). In such cases the user must select areas or elements of the product structure to retrieve relevant information. In contrast to free text search where the query can be modified to manipulate results, when pre-classification is used optimising precision and recall are ever more critical.

2.3 Precision, Recall and F1 Score

The evaluation of IR systems has traditionally involved the measures of *precision*, *recall* and *f1-score* (Witten & Frank 2005). *Precision* is the number of relevant results returned divided by the total number of retrieved results (Equation 4). Maximum precision would be a system that returned every correct result in the corpus and none of the incorrect results. In reality, maximum precision is rarely achieved. *Recall* is then the number of relevant results returned divided by the number of relevant results that should have been returned (*ground truth*) (Equation 5). To obtain a better understanding of an IR systems effectiveness it is important that these two measures be used together and the *f1-score* or *f-measure* combines the two to give a single measure (Equation 6).

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{4}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{5}$$

$$f1\text{–}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

3 RESEARCH APPROACH

This section outlines the overall experimental approach and the implementation using a Formula Student project and specifically a collection of student feasibility and final year reports. The approach comprises of five components: a standard TF-IDF Search Engine, the Product Structure, Example Text, a TF-IDF Model and a Classification Algorithm to classify documents against the model. Each of these is now discussed.

3.1 TF-IDF Search Engine

The TF-IDF search engine generates a matrix that represents each document in the corpus as a list of terms and a corresponding TF-IDF weight that reflects the importance of that term to that document within the context of the corpus. Searching the TF-IDF structure with a search query involves retrieving each document containing a non-zero TF-IDF weight for term(s) contained within the query. The retrieved documents are then ranked based on the weight/summed weights with the highest appearing at the top of the results list.

3.2 Product Structure

A product's Bill of Materials (BOM) is a hierarchical tree of the systems, subsystems and components required to construct it. As an example, a finished car comprises of several subsystems: chassis, power train, drive train, etc. and each of these is then comprised of another list of components: engines contain a cylinder head, cylinder block, camshaft, crankshaft, etc. Typically, this hierarchy is stored as either a tree diagram or an indented list. The hierarchy can also be represented using parent, child and sibling terminology: if a system (e.g. *engine* and *drivetrain*) comprises of subsystems or components (e.g. *engine*, *fuel system*, *exhaust system*, etc.) then the system is the parent and subsystems the children. The children then are all siblings to each other.

3.3 Example Text

Example text should provide a textual 'blueprint' that truly and accurately represents the terminology and style of writing of the documents within the corpus and how they refer to the components within the BOM. From this, a model will be generated and a classification system will attempt to match documents with the most similar body of text and hence the product structure. This body of text should also be of similar length (number of words and sentences) across sibling components (although this is not always possible).

3.4 TF-IDF Model

The TF-IDF model weighting works in the same way as the TF-IDF search engine but at a localised level using example text relevant to the subsystem rather than at a corpus level and using the entire document set. Weights are scored within the context of sibling components rather than the entirety of the corpus, generating a localised weighting. Each component in the BOM is then represented with a unique list of terms and weights that differentiate it from the rest of BOM components.

3.5 Classification Algorithm

Document Classification involves calculating a similarity score between each component in the BOM and each document within the corpus. Approaches like Cosine Similarity (Baeza-Yates et al. 1999; Bird et al. 2009; Witten & Frank 2005) are commonly used by search engines. The approach used is outlined in (Bird et al. 2009), this splits the document into individual terms to form document vectors, calculates and sums the accumulative term TF-IDF scores, and ranks documents based on this score. It is worth noting that both approaches allow each document to relate to more than one component.

4 IMPLEMENTATION

The Institution of Mechanical Engineers Formula Student is a competition that requires University students to design, build and race a single seat racing car. Teams of around 30 students design the car in their third year and construct it in their fourth year. For the purposes of this study, the 2013-14, 2014-15 and 2015-16 reports from the University of Bath were used. Reports are in pdf format, around ten pages in length and comprises of raw unstructured text, tables and images. In total this corpus is comprised of 281 reports. Table 1 shows a more detailed description of the textual content of the corpus.

Table 1. Formula Student Corpus Statistics

Statistic	Number
Documents	281
Average words per document	4671
Average Unique words per document	724

Figure 2 shows the process diagram for creating a traditional TF-IDF search engine, and Figure 2 shows the combined approach. Figure 2 stages 1 and 2 show the construction of the TF-IDF search engine. This was achieved by extracting the text from all 281 reports and generating the TF-IDF weights for each term in the corpus. These were then stored in an index that was interrogated to produce the combined index and the TF-IDF results for the comparison.

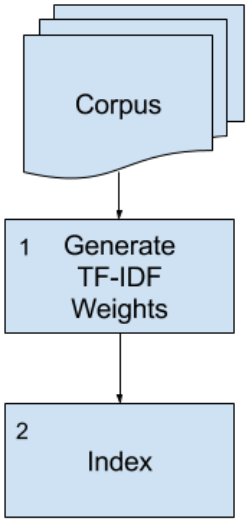

Figure 1. Process diagram for a traditional TF-IDF search index

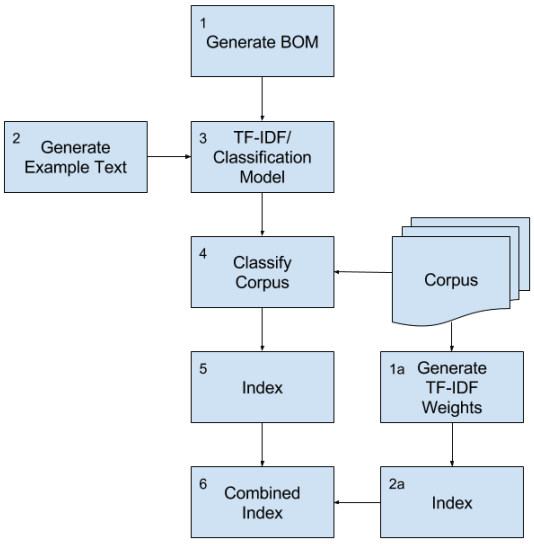

Figure 2. Process diagram for the combined BOM classification

In addition to the reports, students submit a financial statement which includes a BOM. Each year students generate a new design and with it a slightly new BOM. For the purposes of this study a generic BOM was determined using the most common components and component names from across the three years. Table 2 shows an extract from this generic BOM. This relates to stage 1 in the process diagram shown in Figure 1.

Table 2. An extract from the generic BOM

Generic Formula Student BOM			
Level 1	Level 2	Level 3	Level 4
...	...	...	...
fs car	engine and drivetrain	exhaust system	rear exhaust primary
fs car	engine and drivetrain	exhaust system	collector
fs car	engine and drivetrain	exhaust system	secondary pipe
fs car	engine and drivetrain	exhaust system	muffler
fs car	engine and drivetrain	oil system	oil tank deaerator
fs car	engine and drivetrain	oil system	overflow bottle holder
fs car	engine and drivetrain	oil system	overflow bottle
fs car	engine and drivetrain	oil system	oil coolant heat exchanger
fs car	engine and drivetrain	fuel system	fuel tank
...	...	...	...

The next stage (2) involves obtaining component level example text, this was achieved by asking a domain expert to generate bodies of text for each component in the BOM. The domain expert had 30

years' engineering experience with 10 years specifically on Formula Student. An extract from the result is shown in Table 3. On average, each example text contained 31 words split over two sentences.

Table 3. An extract from the BOM and associated example text

BOM Component	Example Text
anti-roll bar	The anti-roll bar is part of the suspension system. Its function is to reduce the body roll of a vehicle during fast cornering or over road irregularities. It connects opposite (left/right) wheels together through short lever arms linked by a torsion spring.
balance bar	A balance bar is an adjustable lever that is pivoted on spherical bearings and employs two individual master cylinders for the actuation of the front and rear brakes. It forms part of the pedal assembly and also provides a mounting for the master cylinders. When the balance bar is cantered, it pushes equally on both master cylinders creating equal pressure.
battery	The battery is an electrochemical device that supplies the electric power to the low voltage system on the vehicle.
bearing	A collective term for is a machine element that constrains relative motion to only the desired motion, and reduces friction between moving parts.
...	...

Stage 3 constructed the model by parsing each component's example text and performing a TF-IDF comparison with its siblings' components example text. TF-IDF scores were then stored for each component. An extract from the model for the component '*Crank Sensor*' is shown in Table 4.

Table 4. An extract from the TF-IDF weight for terms for a BOM component example text

Component	Term	TF-IDF Weight
Crank Sensor	crank	0.0084170581
	speed	0.0084170581
	crankshaft	0.0084170581
	combustion	0.0068013304
	engine	0.0068013304
	internal	0.0068013304
	rotational	0.0058561903
	monitor	0.0058561903
	sensor	0.0051856027

Stage 4 includes the classification of the FS reports. This was done by extracting raw text from the pdf reports and tokenizing the text to obtain a list of terms in each report. For each component in the BOM model the TF-IDF weight for any shared term was summed to give each document a similarity score.

By its nature, the model generates a similarity score for all documents within the corpus. This is counterproductive in IR systems given the system could return every document in the corpus when a search is performed. The only difference between searches is therefore in the order that results are returned. Hence, there is a need to restrict the number results returned by the classification system to those most relevant. The localised (component and system) level weighting of terms within the BOM structure means weight cannot be compared across the entire product structure and so a simple generic threshold score could not be universally implements. The technique used is explored and discussed in the results section.

In addition to the cut off, the Results section also discusses techniques for combining the two sets of results. Essentially, the Results and Discussion and Conclusion section of this paper discuss techniques for delivering an effective Stage 6. Stages 1 and 2 show the construction of the TF-IDF search engine using the exact same method as shown in Figure 2.

A search involves traversing both the TF-IDF search index and BOM Classification search index. The results from both are then combined and ranked based on the summation of the two scores and this is discussed in the following sections. For the purposes of the study and to explore the potential benefits, seven component names were selected at random from the BOM.

5 RESULTS

This section focuses on two main areas. The first is a strategy to limit the number of results returned by the BOM Classification system. The second examines two approaches to combine the results from the two techniques. In order to evaluate the search results from the proposed approach and the traditional TF-IDF, searches were performed on seven terms/components selected at random from the BOM (shown in Table 5). For comparison, the corpus and seven terms were given to a domain expert who generated the ground truth. From the seven terms used, five returned results for both approaches, see Table 5. Neither approach returned 100% of the ground truth documents. The BOM Classification approach returned a higher number of relevant results however, as expected, Table 6 shows how the approach also returns a far higher number of non-relevant results.

Table 5. The number of retrieved relevant documents returned by each approach

Query	Ground Truth	Retrieved Relevant Documents	
		BOM Classification	TF-IDF
Brake System	38	22	19
Cooling system	21	17	13
Exhaust System	40	30	19
Front Wing Assembly	26	21	7
Paint - Body	2	0	0
Steering Column	8	6	5
Track Rod	3	0	1

Table 6. The total number of retrieved documents returned by each approach

Query	Ground Truth	Total Retrieved Documents	
		BOM Classification	TF-IDF
Brake System	38	229	86
Cooling system	21	232	81
Exhaust System	40	240	66
Front Wing Assembly	26	238	45
Paint - Body	2	0	0
Steering Column	8	238	22
Track Rod	3	0	48

5.1 Cut-Off

Several techniques were tested to find a representative cut-off for the classification results, top n-results or top x-percent, for example. Further study in this area is needed and so this paper does not focus on this investigation. However, the closest representative measure found was using the number of results returned by the TF-IDF approach.

Figure 3 and Figure 4 show the number of relevant documents returned versus the total number of documents returned for the '*Brake System*' and '*Exhaust System*' respectfully. These two figures are included because they are examples of the two trends that were seen in the results. The *Brake System* (Figure 3) shows how the TF-IDF line generates an appropriate cut-off point for the BOM Classification results as most relevant results are distributed in the first 50-100 results. However, the *Exhaust System* (Figure 4) shows how the BOM Classification approach distributes relevant results across the total number of documents retrieved. Of the five terms with results for both approaches, three follow a similar pattern to the *Brake System* while the other two follow the *Exhaust System*. Given a cut-off is needed, the total number of result returned by TF-IDF was chosen.

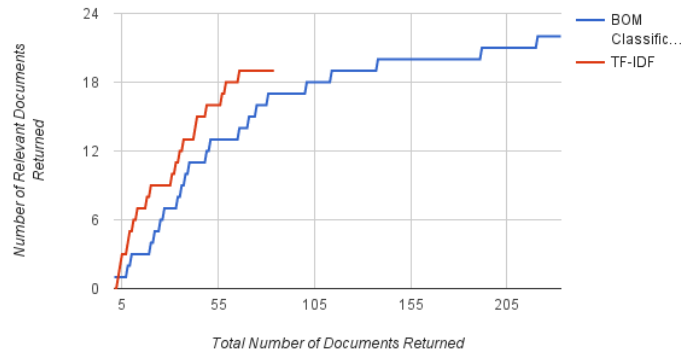

Figure 3. Brake System: Relevant versus Total Returned Results

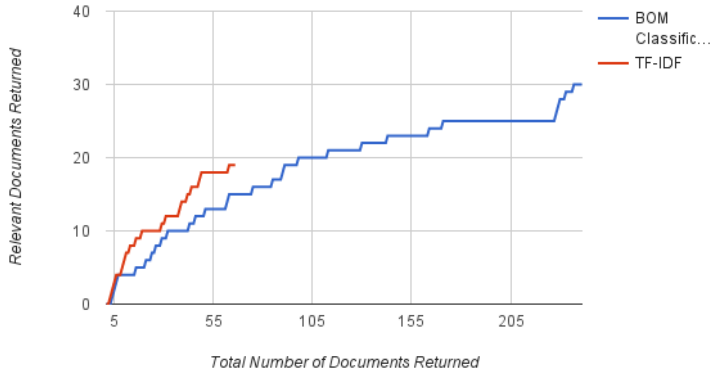

Figure 4. Exhaust System: Relevant versus Total Returned Results

5.2 Combining Results

The Precision, Recall and F1 Score for the merger (Figure 5 and Figure 7) and intersection (Figure 6 and Figure 8) of the two sets of results after the cut-off is shown below. Merging results involves combining the results in order one at a time before removing the duplicates. When removing duplicates the highest ranks results was kept. The intersect filters only those results that appear in both approaches and again the highest rank for each result was kept.

The first noticeable difference is how the merged results boost the recall while reducing the Precision and F1-Score. The opposite is true of the intersect where the Precision and F1-Score are increased while a drop in the Recall occurs for two terms and no change is seen in the other five terms.

Both figures do however show a result that goes against these trends. The large percentage improvement for the '*Front Wing Assembly*' causes a very slight increase in the F1-Score. Looking at the figure for the intersect, a large percentage decrease in the '*Exhaust System*' also causes a decrease in the F1-Score.

Figure 7 and Figure 8 show the same data as Figure 3 and Figure 4 with the addition of the values for the merged and intersected results. Again, these two examples are typical of those seen across the five terms that returned results for both approaches. Both figures show how the results for the merged results lie between the TF-IDF and BOM Classification results - as one would expect. The line reaches a maximum at a point that is either equal to or exceeding the TF-IDF approach but does so over a larger number of returned results. Both figures also show how the intersection of the two approaches delivers relevant results sooner - with fewer non-relevant results returned, this is at the cost of the number of relevant results returned with the line stopping short of the lines for the other three approaches.

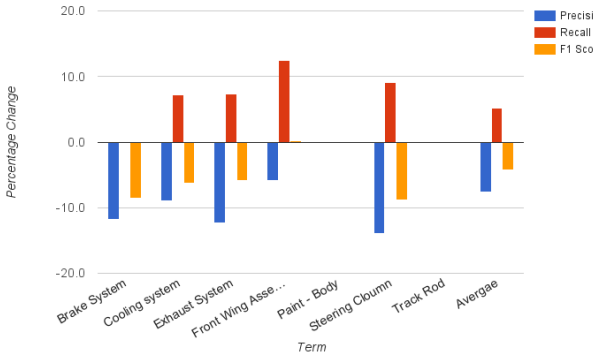

Figure 5. Precision, Recall and F1 Score for the merged results

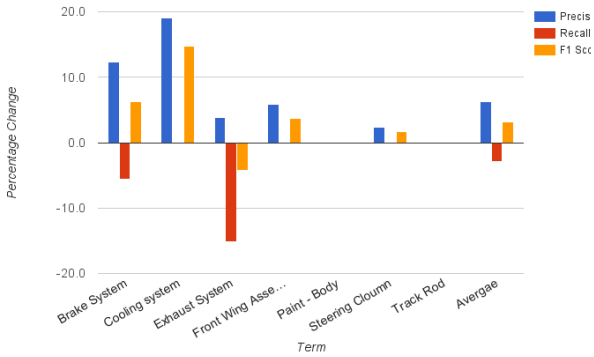

Figure 6. Precision, Recall and F1 Score for the Intersected results

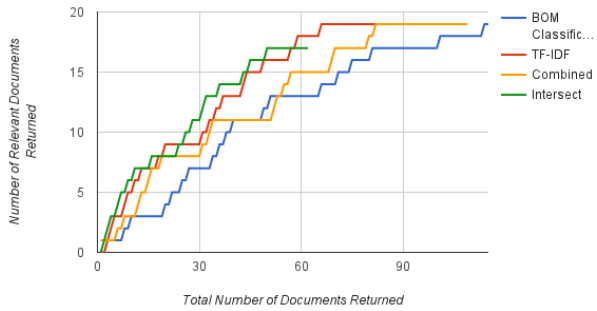

Figure 7. Precision, Recall and F1 Score for the merged results

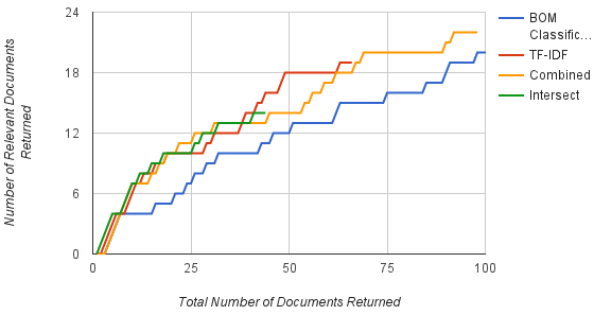

Figure 8. Precision, Recall and F1 Score for the Intersected results

6 CONCLUSION AND DISCUSSION

The approach outlined in this paper aimed to improve a TF-IDF search system by combining the TF-IDF with a classification system that classified documents against the product structure. To investigate this concept a search engine was constructed that generated results for the two approaches and methods for combining these results were examined. The results show that the number of results returned by a TF-IDF search generated a representative cut-off point for the number of results returned by the BOM Classification approach used. In combining the results from the two approaches, promise was seen depending on whether the goal is to expand on the number of relevant results returned or to increase the precision of those results.

The goal of any search engine is 100% precision but challenges such as the ambiguity of language and the uncertainty in the user's information needs mean that there are no perfect IR systems and all searches are carried out with the expectation that the final stages of the search will be performed by our own evaluation and browsing of a corpus subset. The work presented here shows that the technique provides some tailoring of this subset. An example its usefulness was presented by the authors in (Jones, Chanchevrier, et al. 2015) where IR is performed via a three-dimensional visual representation of the artefact. Representing many results within a three-dimensional artefact space will quickly swap the visualisation and become unusable. In this case, the intersect of the two result sets will benefit the visualisation, and the possible reduction in the recall may be acceptable given an increase in precision and greatly reduced number of documents returned.

There are however several areas that warrant further research. The size of the corpus used here (281 documents) did not allow for the more traditional division of the corpus into training and testing sets and so example text was generated and used. One can question whether the example text is representative of the corpus itself and it would be beneficial to repeat this study on a larger corpus, for example those used within large organisations such as Aerospace where mature products generate larger corpora, standardised lexicons, reporting/documentations and procedures. It is worth noting, however, that the corpus used is a 'real world' example and while results may improve with a larger dataset (thousands or

tens of thousands of documents), research will at some point need to deliver solutions for these 'real world' challenges of small corpora.
There are several alternative machine learning approaches to construct the classification model, for example artificial neural networks and deep learning. The aim of the work presented here was to determine if the product structure can be used to improve IR and not to determine the best approach for doing so. Now that it has been shown that the technique can have a positive impact on the results returned the foundations are in place for this future work.
The final aspect to discuss is whether there are better strategies for merging the two results sets. This study showed the intersect improved precision while the merging improved the Recall and F1-Score. No attempt was made to integrate the two approaches and generate a result set that optimised all three measures. This would also benefit from further study.

REFERENCES

Baeza-Yates, R., Ribeiro-Neto, B. & others, (1999), *Modern information retrieval*, ACM press New York.

Bird, S., Klein, E. & Loper, E., (2009), *Natural language processing with Python*, " O'Reilly Media, Inc."

Camba, J. et al., (2014), "Extended 3D annotations as a new mechanism to explicitly communicate geometric design intent and increase CAD model reusability". *Computer-Aided Design*, 57, pp.61–73.

Chen, H., Chiang, R.H.L. & Storey, V.C., (2012), *Business Intelligence and Analytics: From Big Data to Big Impact*, [online] Available at: http://aisel.aisnet.org/misq/vol36/iss4/16

Eastman, C. et al., (2011), BIM handbook: *A guide to building information modeling for owners, managers, designers, engineers and contractors*, John Wiley & Sons.

Finkelstein, L. et al., (2001), "Placing search in context: The concept revisited", *Proceedings of the 10th international conference on World Wide Web*. pp. 406–414.

Hawking, D., (2004), "Challenges in enterprise search", *Proceedings of the 15th Australasian database conference*, Volume 27. pp. 15–24.

Hicks, B.J. et al., (2002), "A framework for the requirements of capturing, storing and reusing information and knowledge in engineering design", *International Journal of Information Management*, 22(4), pp.263–280.

Jones, D.E., Chanchevrier, N., et al., (2015), "A STRATEGY FOR ARTEFACT-BASED INFORMATION NAVIGATION IN LARGE ENGINEERING ORGANISATIONS", *DS 80-10 Proceedings of the 20th International Conference on Engineering Design (ICED 15)*, Vol 10: Design Information and Knowledge Management Milan, Italy, 27-30.07. 15

Jones, D.E., Xie, Y., et al., (2015), "Improving Enterprise Wide Search in Large Engineering Multinationals: A Linguistic Comparison of the Structures of Internet-Search and Enterprise-Search Queries", *IFIP International Conference on Product Lifecycle Management*, pp. 216–226.

Klampanos, I.A., (2009), (Ed.) Manning Christopher, Prabhakar Raghavan, Hinrich Schütze: "Introduction to information retrieval", In: *Information Retrieval*, 12(5), pp.609–612. [online] Available at: http://link.springer.com/10.1007/s10791-009-9096-x

Lee, S.G. et al., (2008), "Product lifecycle management in aviation maintenance, repair and overhaul", *Computers in industry*, 59(2), pp.296–303.

Liu, D.T. & Xu, X.W., (2001), "A review of web-based product data management systems", *Computers in industry*, 44(3), pp.251–262.

Stocker, A. et al., (2014), "Is enterprise search useful at all?: lessons learned from studying user behavior", *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, p. 22.

Witten, I.H. & Frank, E., (2005), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Xie, Y. et al., (2011), "Applying context to organize unstructured information in aerospace industry", *DS 68-6: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design*, Vol. 6: Design Information and Knowledge, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011

ACKNOWLEDGMENTS

This research is funded via an EPSRC CASE AWARD, the Language of Collaborative Manufacturing (LOCM) Project (EPSRC grant reference EP/K014196/1) and the Airbus Group. The authors would like to thank colleagues at the Airbus Group, University of Bristol, University of West England and the University of Bath for their support and contribution.